Discrete Applied Mathematics 157 (2009) 2378–2387

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

Prokaryote clustering based on DNA curvature distributions

L. Kozobay-Avraham [a,b], S. Hosid [a,b], Z. Volkovich [c], A. Bolshoy [a,b,*]

[a] Department of Evolutionary and Environmental Biology, University of Haifa, Haifa 39105, Israel

[b] Genome Diversity Center of Institute of Evolution, University of Haifa, Haifa 39105, Israel

[c] Software Engineering Department, ORT Braude College of Engineering, Karmiel 21982, Israel

ARTICLE INFO

Article history:
Received 22 January 2007
Received in revised form 20 June 2007
Accepted 11 June 2008
Available online 23 September 2008

Keywords:
Curved DNA
Clustering methods
k-means
PAM

ABSTRACT

Massive determination of complete genome sequences has led to the development of different tools for genome comparisons. Our approach is to compare genomes according to typical genomic distributions of a mathematical function that reflects a certain biological function. In this study we used comprehensive genome analysis of DNA curvature distributions in coding and non-coding regions of prokaryotic genomes to evaluate the assistance of mathematical and statistical procedures. Due to an extensive amount of data we were able to define the factors influencing the curvature distribution in promoter and terminator regions such as growth temperature, genome size, and $A+T$ composition. Two clustering methods, K-means and PAM, were applied and produced very similar clusterings that reflect genomic attributes and environmental conditions of the species' habitat.

1. Introduction

The term DNA curvature refers to a characteristic of DNA fragments, which are bent without application of any external forces. This property is also called intrinsic curvature or sequence-dependent DNA curvature. The presence of curved DNA was established by biological experiments in the early 1980s (see, [22,35,4,7,31]). Based on experimental results, some computational models, including our model [2,26], were developed to predict the magnitude of DNA curvature with high reliability. The existence of upstream curved sequences (UCS) was shown experimentally for many genes in prokaryotes [25]. Comprehensive genome analysis of DNA curvature in regulatory regions was performed by us [3,17,16,18], and others [12, 23].

It is well known that the genomes are annotated with quite dramatically varying degrees of quality. The most striking examples are the *A.* pernix and *P.* horikoshii, which are two of a small handful of genomes that are classified as hyperthermophiles. Naturally, one could suspect that part of the trends that might be seen is simply artifactual fluctuations, due to differing qualities in gene finding, rather than "real" differences. Our analysis brought us to the conclusion there is no systematic bias in the quality of Archaeal gene predictions (data not shown).

Offered approach was to detect predicted frequent occurrences of especially high local extremes of curvature distribution function upstream to starts of the coding sequences. Following detection of similarly outstanding curvature distribution downstream of genes in *Escherichia coli* and *Bacillus subtilis*, wide genomic comparisons (170 complete prokaryotic genomes) were performed, and we found that not only upstream, but also downstream, intergenic regions are significantly more curved than would be expected from their dinucleotide composition [18]. Putative influence of environmental and genomic factors as well as taxonomic factors on curvature distribution in promoter and terminator regions were indicated [3,17,16,

* Corresponding author. Tel.: +972 4 8240382; fax: +972 4 8240382.
E-mail address: bolshoy@research.haifa.ac.il (A. Bolshoy).

0166-218X/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.dam.2008.06.049

18]. The most prominent effect on DNA curvature distribution, in these regulatory regions, was presented by the growth temperature.

To avoid any misunderstanding we want to mention that everywhere in this manuscript the term "start of gene" means "start of translation or a position of a first codon", and "end of gene" means "end of translation or a position of a stop codon". "Start of gene" does not refer to the transcription +1 site! The reason for this usage is rather simple: transcription initiation site is not known for most of the more than 200 genomes used.

Cluster analysis and other statistical tests were performed on genomic curvature distributions data. For every genome, we predicted DNA curvature profiles: one typical profile for the start of translation (covering a promoter region and a beginning of a coding region) and another one around the end of genes (3′-end of coding sequence and a putative transcription terminator region). The profiles were predicted using the CURVATURE program [26]. Actually, instead of using raw curvature distributions for clustering, we used normalized data. Randomized sequences were constructed and curvature excess profiles in standard deviation units were calculated for each genome.

The normalization distinguishes our attitude to construction curvature profiles from the profiles used in other studies [24, 33,3].

Six different distances have been examined for the purpose of future clustering. The squared Euclidean distance between the genomes appears to be the most reasonable from a biological point of view. Consequently, this distance has been used for all further mentioned results of clustering.

2. Methods

2.1. Clustering methods

Clustering problems arise in many areas of bioinformatics. Clustering is an example of unsupervised learning when the number and type of classes are unknown, and available data samples are unlabeled. Groups (clusters) are constructed to achieve a relatively high similarity among the groups' elements in addition to a relatively low similarity between elements of different groups. Let us consider a subset $X = \{x_1, x_2, \ldots, x_m\}$ in the n-dimensional Euclidean space R^n. Consider a partition $\Pi = \{\pi_1, \ldots, \pi_k\}$ of the set, i.e.

$$\bigcup_{j=1}^{k} \pi_j = X, \qquad \pi_i \cap \pi_j = \varnothing \quad \text{for } i \neq j. \tag{1}$$

For a real-valued function q whose domain is the set of subsets of X the quality of the partition is defined as

$$Q\,(\mathbf{\Pi}) = \sum_{j=1}^{k} q(\pi_j). \tag{2}$$

In fact, clustering problem is another instance of a global optimization problem of finding a partition

$$\Pi^{(0)} = \{\pi_j^{(0)}, j = 1, \ldots, k\}, \tag{3}$$

which optimizes $Q(\Pi)$. Often function q is associated with a "dissimilarity measure", or a distance-like function $d(x, y)$. The term a distance-like function is used, since this function is not required to satisfy all requirements to a distance function: the function is not necessarily symmetrical or necessarily satisfies the triangle inequality. Function q can be constructed by means of $d(x, y)$ as such. We introduce k centroids (medoids) $(c_1, \ldots, c_k)$ as a prescribed subset of R^n. This set defines a partition of X as

$$\pi_i = \{x \in X : d(c_i, x) \leq d(c_j, x), \text{ for } i \neq j\}. \tag{4}$$

(Ties are broken arbitrarily.) On the other hand, for a given partition the centroids set is defined as

$$c(\pi_i) = \arg\min_{c_i} \left\{ \sum_{x \in \pi_i} d(c_i, x) \right\}. \tag{5}$$

Thus,

$$q(\pi_i) = \sum_{x \in \pi_i} d(c(\pi_i), x) \tag{6}$$

and the mentioned optimization problem is reduced to finding an appropriate centroids' set as a solution to the task

$$C = \arg\min_{c_1, \ldots, c_k} \left\{ \sum_{i=1}^{k} \sum_{x \in \pi_i} d(c_i, x) \right\}. \tag{7}$$

Dissimilarity measures can be chosen in different ways, particularly as information distances of divergences (see, for example [28,15]). We consider the following six cases for two vectors $x = \{x_i, i = 1, \ldots, n\}$ and $y = \{y_i, i = 1, \ldots, n\}$:

(1) The squared Euclidean distance $d(x, y) = \|x - y\|^2$;
(2) The Manhattan distance $d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$;
(3) The max distance $d(x, y) = \max_i |x_i - y_i|$;
(4) The correlation distances $d(x, y) = 1 - r(x, y)$, where $r(x, y)$ is one of the following correlation coefficients:
   - The correlation coefficient of Spearman;
   - The rank correlation coefficient of Pearson;
   - The rank correlation coefficient of Kendall.

Apparently, the most widespread iterative procedures for an approximate solution of the clustering (5) are the *k*-means and the PAM (Partition Around Medoids) algorithms. The PAM algorithm [13] has the pairwise distance matrix and the suggested number of clusters *k* as input parameters. The algorithm was developed to find the *k* most representative objects (medoids) that represent *k* clusters such that non-selected objects are clustered with the medoid to which it is the most similar. The total distance between non-medoid objects and their representative medoid may be reduced by swapping one of the medoids with one of the objects iteratively. Obviously, it is time consuming even for the moderate number of objects and small number of medoids. The PAM approach looks more robust and efficient than the *k*-means algorithm and is implemented in clustering packages *R* and *S − Plus*.

The classical *k*-means algorithm [5] is based on the squared Euclidean distance, however many generalizations can be found in the literature (see, for example, [28,15]). The *k*-means algorithm is being considered as a simplification of the well known Expectation Maximization (*EM*) algorithm (see, for example [6]) and corresponds to the so-called spherical clusters case. On one hand, this assumption is hardly ever satisfied in practice; as a consequence, the attained solution does not necessarily match the most optimal solution, which corresponds to the global objective function minimum. As a result, so-called "non-optimal stable clusters" can be created. On the other hand, it can be demonstrated that the procedure always converges. The provided numerical experiments show that in our situation the *k*-means produces clusters having preferred biological meaning. Due to this reason the algorithm was employed. The algorithm is composed of the following steps:

(1) Set *k* points into the space represented by the items that are being clustered. These points stand for initial cluster centroids.
(2) Classification: assign each item to the closest centroid.
(3) Minimization: calculate the mean (centroid) of each cluster.
(4) Repeat Steps 2 and 3 until the partition is stable, that is, until the centroids no longer change.

The following cluster stability indexes are used frequently. Given a partition $\mathbf{\Pi} = \{\pi_j, j = 1, \ldots, k\},\ 2 \le k$, we denote by B_k and W_k the dispersion matrices of between and within group sums of squares (see, for example, [21]).

(1) The Krzanowski and Lai index [19] is defined by the following relationships:

$$\mathrm{diff}_k = (k-1)^{\frac{2}{n}} \mathrm{tr}(W_{k-1}) - k^{\frac{2}{n}} \mathrm{tr}(W_k) \tag{8}$$

$$KL_k = \left| \frac{\mathrm{diff}_k}{\mathrm{diff}_{k+1}} \right|. \tag{9}$$

The estimated number of clusters is the maximal value of the index KL_k.

(2) Sugar and James [27] proposed an information theoretic approach for an estimation of the true number of clusters. A transformation power t is predetermined (a typical value is $t = n/2$).

$$J_k = \mathrm{tr}(W_k)^{-t} - \mathrm{tr}(W_{k-1})^{-t}.$$

The estimated number of clusters maximizes the value of the index J_k.

2.2. Biological methods

2.2.1. Genomic sequences and their attributes

For further analysis we took 205 complete prokaryotic genomes from the GenBank, a public genome library of the National Center for Biotechnology Information. The following genomic characteristics were gathered from the genomic annotations and from the literature: optimal growth temperature, genome size, $A + T$ composition, and taxonomic description.

- Optimal growth temperature — the organisms belong to four temperature groups, as defined in literature:
- psychophiles — organisms that are defined by their ability to grow at 0 °C and below;
- mesophiles — organisms that grow best between 10 °C and 30 °C;
- thermophiles — organisms that grow best in hot conditions, between 30 °C and 50 °C;
- hyperthermophiles — organisms that thrive in extremely hot environments, i.e., hotter than 60 °C with optimal temperatures between 80 °C and 110 °C.

- Taxonomy — prokaryotes fall into one of two groups, Archaebacteria (ancient forms thought to have evolved separately from other bacteria) and Eubacteria. Archaebacteria, sometimes called Archaea, emerged at least 3.5 billion years ago and lived in environments that existed when the earth was young. Many hyperthermophiles are Archaea. Our genomic database consists of 23 Archaeal and 182 Bacterial genomes. Among the Archaeal genomes 19 representatives are thermophiles or hyperthermophiles; four genomes are mesophiles.
- Genome sizes — prokaryotes have relatively small genomes: from very short genomes with lengths less than 1Mb - up to about 9Mb. The size of a genome is a relevant factor because the smallest genome-sized prokaryotic species, the obligate endocellular parasites, when compared to their free-living relatives, have preferentially lost many regulatory elements, including factors. This phenomenon is probably due to the rather stable environment inside host cells, which renders extensive gene regulation useless [1,32,14]. DNA curvature excess is related to gene regulation [3,17,18].
- $A + T$ composition — this feature is relevant to our analysis because the magnitude of DNA curvature is related to it. In AT-rich segments curved DNA fragments occur more frequently than in AT-poor segments. Moreover, it was shown that strongly curved DNA fragments must possess high $A + T$ content.

2.2.2. Curvature calculation

At a point P, the osculating circle is the best circle that approximates the curve at P. Ignoring degenerate curves such as straight lines, the osculating circle of a given curve at a given point is unique. Curvature K at a point P is calculated from the formula

$$R = \frac{1}{|K|},$$

where R is the radius of the osculating circle at the point P. As we mentioned in our manuscript [26] this, classical curvature gives us an inappropriate measure of DNA intrinsic curvature. We proposed to use a measure dependent on an arc size. Instead of taking the best circle that approximates the curve at P we proposed to use the best circle approximating a path segment of the length equal to a program parameter arc size. Our CURVATURE algorithm [26] is based on the stepwise calculation of geometric transformations according to the set of previously estimated parameters [2]. Firstly, in our algorithm we calculated DNA path. Than calculations of a DNA map follow. At every position the best approximating arc may be defined. A curvature value at a sequence-position P corresponds to a curvature of the arc approximating the predicted DNA path, where the arc approximates a path segment of the length equal to a program parameter arc size with a center of the segment in the position P. As can readily be seen from the definition, curvature therefore has units of inverse distance. Traditionally, we use normalized DNA curvature units (cu) introduced by Trifonov and Ulanovsky [30]. One curvature unit corresponds to the mean DNA curvature in the crystallized nucleosome ($1/42.8A^*$).

$$\kappa = \frac{K}{K_{\text{nucleosome}}} = \frac{42.8A^*}{\text{radius of the approximating arc}}. \tag{10}$$

For example, a segment of 125 bp of length with a shape close to a half–circle has a curvature value of about 0.34 cu. Such strongly curved pieces appear infrequently in genomic sequences. The program is available upon request from A. Bolshoy at bolshoy@research.haifa.ac.il.

2.2.3. Construction of curvature-excess profile

The whole genome sequence was used as input for the CURVATURE program and a map of curvature distribution using a given window size of 125 base pairs (bp) along the whole sequences was produced. To construct genomic profiles before the start of genes, after the start of genes, or after the end of genes relevant pieces of produced DNA curvature genome maps were averaged.

- Averaging — gene positions were taken from genome annotations that accompany every genome sequence in the GenBank. In our study of curvature distribution around the starts or ends of genes, we only processed genes flanked by intergenic regions longer than 125 nucleotides. The reason for this choice is that shorter intergenic regions can hardly include regulatory signals. Hundreds of genes of a genome i were processed to obtain a genomic profile (g_i). The standard errors (s_i) were estimated by bootstrap method using 1000 runs.
- Preparation of randomized genomes — we constructed control genomes for testing the significance of the results and comparing properties of natural and artificial genomes. The construction procedure consisted of three steps:
- a genome was cut in separate genic and intergenic pieces at every start and end gene junction;
- each piece was separately reshuffled preserving dinucleotide composition;
- all the pieces were reassembled in the original order.

  A randomized DNA curvature map was obtained by averaging maps of ten shuffled genomes. To construct randomized profiles relevant pieces of produced randomized genome maps were averaged (r_i).
- Curvature excess calculation — curvature excess related to a genome i is an apparent deviation between genomic profile (g_i) and randomized profile (r_i). Excess curvature value at position k is measured in standard deviation units (s_i) and calculated as follows:

$$ce_{ik} = (g_{ik} - r_{ik})/(s_{ik}). \tag{11}$$

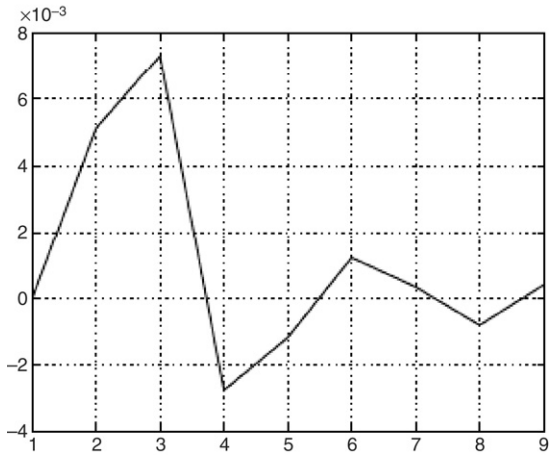

Fig. 1. Graph of the Sugar and James index.

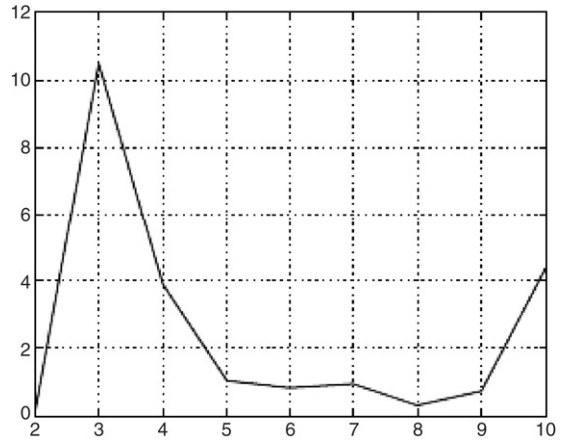

Fig. 2. Graph of the Krzanowski and Lai index.

3. Results

3.1. Preliminary study of clustering analyses

In order to evaluate which of the known distance calculations is the most suitable for our data, the following distances were calculated between curvature excess profiles in promoter regions: the squared Euclidean distance, the Manhattan distance, the max distance, and three correlation distances of Spearman, Pearson, and Kendall (see, Section 2). Partitions have been provided by means of the PAM algorithm using all these distributions and by means of the *k*-means algorithm using the squared Euclidean distance. The results clearly point to the last approach as the most appropriate for our data. Therefore, all further cluster analyses in the non-coding regions were performed using this method.

3.2. Clustering of non-coding regions

We performed cluster analyses using the *k*-means and the PAM methods, based on the Euclidean distances between curvature excess profiles before and after the genes.

The graphs (Figs. 1 and 2) of the two above mentioned cluster stability indexes (Clustering methods section) demonstrate that both indexes precisely indicate the true number of clusters as three. The possible number of clusters is presented on the *X*-axis. On the *Y*-axis the values of an appropriate index are plotted.

In all further text the number of clusters is assumed to be three.

3.3. Correlation between clusters and genomic characteristics in promoter regions

Three clusters were obtained using *k*-means algorithm and the graphs present centroid profiles related to each of three clusters. The highest profile is related to cluster 1, and the lowest profile corresponds to cluster 3.

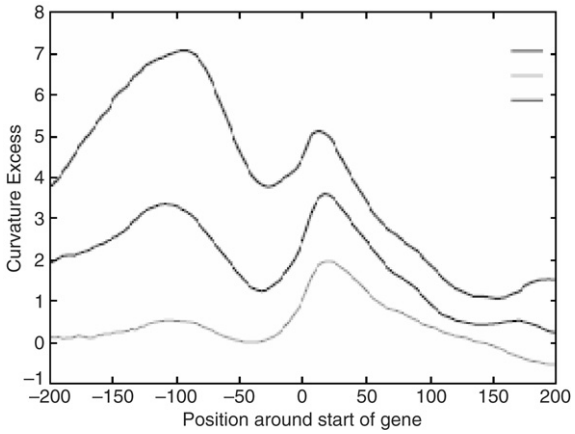

Fig. 3. Genomic profiles based on curvature excess distributions in the neighborhood of the starts of genes. Three clusters were obtained using K-means algorithms and the graphs present centroid profiles related to each of three clusters. The highest profile is related to cluster 1, and the lowest profile corresponds to cluster 3.

Table 1
Cross-tabulation count between temperature and clusters

Temperature/Cluster	Psychophiles	Mesophiles	Thermophiles	Hyperthermophiles
1	0	48	0	0
2	4	67	4	4
3	0	57	8	13

Fig. 3 presents the three genomic profiles according to the clusters obtained by the k-means method based on curvature excess distribution in the neighborhood of the starts of genes. Cluster 1, the smallest cluster containing genomes with the highest curvature excess values in promoter regions, is rather homogeneous. The cluster contains exclusively mesophilic prokaryotes that have genome sizes larger than 1.4 Mb and a high $A+T$ composition. The PAM method gives rather similar results; 92% genomes in the smallest cluster are also big AT-rich mesophilic genomes.

Table 1 presents the results of clustering obtained by the k-means algorithm applied to curvature excess in promoter regions cross-tabulated with temperature classifications. The FM correlation coefficient is equal to 0.48 for the correlation between the temperature and curvature classifications. Cluster 1, as we have mentioned above, contains only mesophilic prokaryotes. One can also learn from this table that a majority of the thermophiles and hyperthermophiles is located in cluster 3. As we can see in Fig. 3 the centroid of this cluster presents the lowest curvature excess level among all clusters. Interestingly, all four psychophiles are located in cluster 2, which contains the genomes with the intermediate curvature excess value in promoter regions. Mesophilic genomes have notable representation in clusters 2 and 3.

The majority of the processed Archaea is hyperthermophiles, and hyperthermophiles are mainly Archaea. Naturally, the correlation between taxonomy and k-means clustering based on curvature promoter profiles is similar to the above mentioned correlation with growth temperatures. However, we observe that clustering of mesophilic Archaea is similar to clustering of mesophilic Eubacteria, while clustering of hyperthermophilic Eubacteria is as of Archaebacteria.

Genome size and $A+T$ composition were also found to have influence on curvature distribution [17]. Lengths of the currently available prokaryotic genomes range in size from 490,885 to 9105,828 bp. In previous publications [17,18] we arbitrarily divided the genomes into two groups with a threshold of 1.4 Mbp. In order to verify our intuitively chosen threshold of 1400,000 bp, t-tests were performed: once according to the median (2.4 Mbp) and second according to the arbitrary value (1.4 Mbp). We found that the differences between the mean values of the groups were significantly higher when the threshold of 1.4 Mbp was used. These results verified our intuitive threshold between 'small' and 'big' genomes used in previous publications [17,18].

With respect to $A+T$ composition, we divided the genomes into four groups:

- "very GC-rich" (0%–30%);
- "GC-rich" (30%–47%);
- "AT-rich" (47%–65%);
- "very AT-rich" (65%–100%).

Here the $A+T$ composition of the genome is a mean value averaged over the complete genome. However, $A+T$ composition of prokaryotic intergenic regions is higher than the appropriate genic sequences in almost all prokaryotic genomes. Therefore, the threshold of 47% between the second and the third AT-rich groups was chosen to include genomes with $A+T$ composition over 50% in intergenic regions. Qualitative correlation between the attained partitions and $A+T$ composition was also observed. Cluster marked as 1 does not contain any GC-rich genomes ($A+T$ composition below 47%).

Table 2
Cross-tabulation count between clusters based on comparison of distributions related to starts and ends of genes

e_clusters[a] / *s_clusters*[b]	1	2	3
1	35	7	0
2	13	50	31
3	0	22	47

[a] Clusters obtained using upstream curvature profiles.
[b] Clusters obtained using curvature distributions after ends of genes.

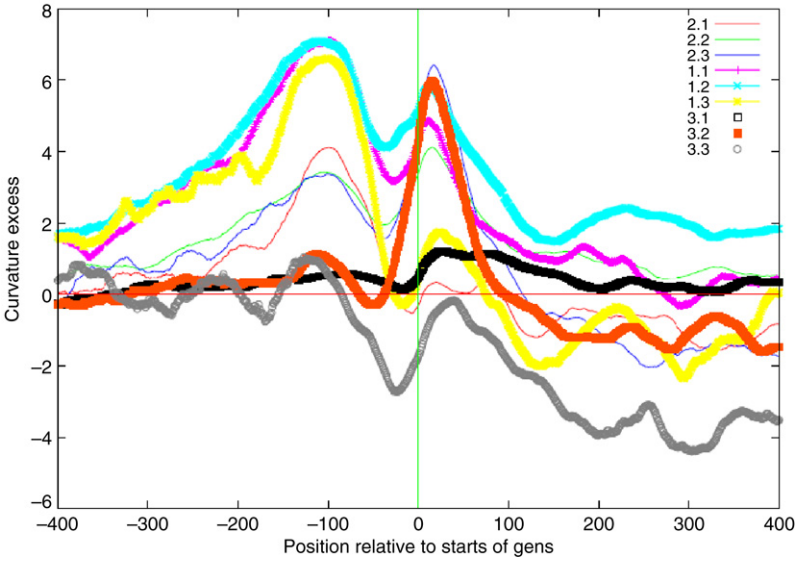

Fig. 4. Genomic profiles in the neighborhood of the starts of genes.

This cluster mainly consists of genomes with $A + T$ content in a range of 47%–65%. Surprisingly, only three genomes (out of 35) from the fourth group, genomes with very high $A + T$ content, were represented in this cluster. Examination of the remaining 32 "very *AT*-rich" genomes revealed that most of them were "small" genomes.

3.4. Correlation between DNA curvature in the neighborhoods of starts and ends of genes

Cluster analysis was also performed on curvature excess distribution immediately after the ends of genes (terminator sites). The correlation coefficient FM is 0.5.

Table 2 presents the cross-tabulation between *s_clusters* (clusters obtained using promoter curvature profiles) and *e_clusters* (clusters obtained using curvature distributions after ends of genes). The most striking features of the table are the high values on the main diagonal; zeros and small values in the other cells related to the clusters with high curvature excess.

3.5. Sub-clustering of coding regions

Centroids of the three clusters in the neighborhood of the starts of genes (from −200 bases to +200 bases) are shown in Fig. 3. The results of the clustering were based on measuring squared Euclidean distance $d_{i,j}$ between profiles x_i and x_j in the upstream region only:

$$d(x, y) = \|x - y\|^2 \;=>\; d_{i,j} = \sum_{l=-400}^{0} (x_{\mathbf{i}}^l - y_{\mathbf{i}}^l)^{\mathbf{2}}.$$

As a further step we considered sub-partitions for each one of the obtained clusters separately in the purpose to achieve more detailed biological interpretation. The *k*-means algorithm accompanied by the Euclidean distance with the number of clusters equal to 3 has been used. The attained sub-partitions can be found in the supplementary. The centroids of all nine clusters are presented in Fig. 4.

Several surprising features of curvature excess distributions in the coding regions were observed. One of them is a sharp maximum immediately after the start of translation typical for sub-clusters 1,1 and 1,2; 2,2 and 2,3; and 3,2. Let us denote this group as "Genic 5-end Maximum Characterized Clusters" or G5MCC. The most surprising among these profiles

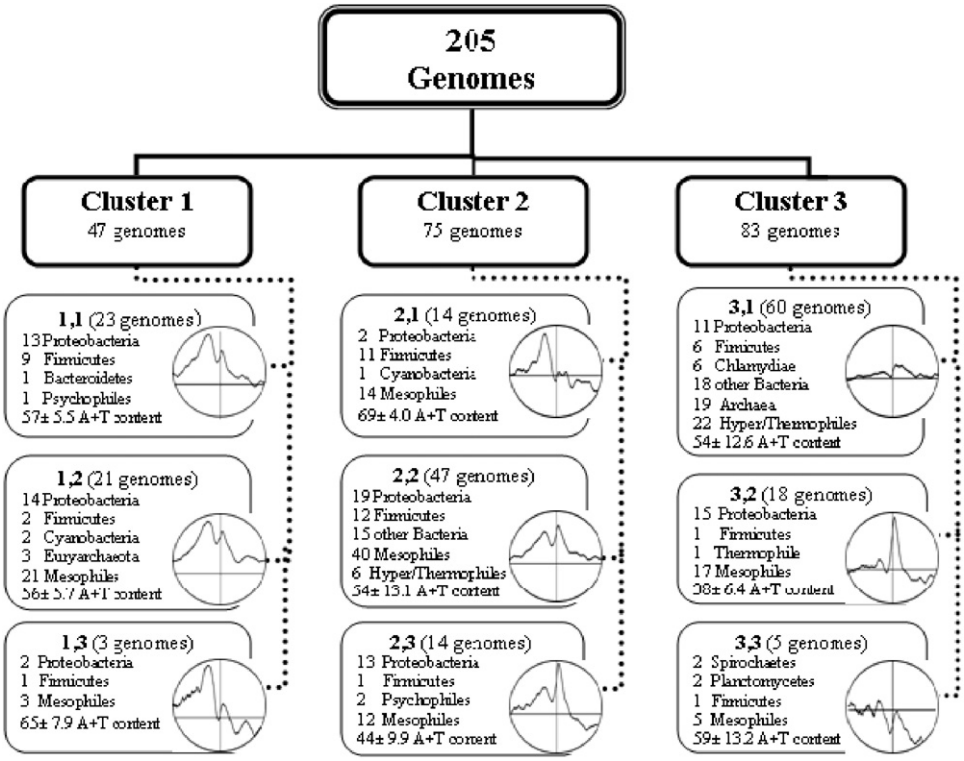

Fig. 5. A schema of clusters and subclusters.

is the sub-cluster 3,2, which belongs to cluster number 3 with the lowest curvature excess values in promoter regions. Sub-cluster 3,2 presents the sharpest inclination of curvature excess profile immediately after the start of genes. Trying to define genomic attributes of G5MCC, we considered environmental, taxonomic, and compositional genomic factors. As we mentioned in our previous publications [3,17], growth temperature and curvature distribution correlate well. We took notice (Fig. 5) that G5MCC are usually mesophiles: all thermophiles (including hyperthermophiles) are located in two of nine clusters — 22 thermophiles belong to the cluster 3,1, which is not G5MCC, but 6 thermophiles also belong to cluster 2,2 (*Geobacillus kaustophilus, Streptococcus thermophilus, Thermoanaerobacter tengcongensis* and *Thermobifida fusca* — thermophiles; *Carboxydothermus hydrogenoformans* and *Pyrococcus horikoshii* — hyperthermophiles).

The profile of the sub-cluster 3 of cluster 1 with the highest curvature excess values in promoter regions has an opposite surprising feature: the sharp descent after the start of genes. However, this group contained only three genomes; therefore, to explain this phenomenon further investigation is needed.

Three clusters were obtained by the *k*-means method based on the Euclidean distances between curvature excess distributions upstream of the starts of genes (see Fig. 3). Next, we considered sub-partitions for each one of the obtained clusters separately using *k*-means algorithm based on curvature excess distributions downstream of the starts of genes. The graph presents centroid profiles related to each of the nine subclusters.

We provided cluster validation tests based on the two above mentioned indexes: Krzanowski and Lai index and Sugar and James index. The results indicate the preferred number of clusters to be 2 or 6.

Clustering was performed as described in the legend to Fig. 4. Cluster notation is as it appeared in Figs. 3 and 4, respectively. An amount of the genomes related to a cluster is indicated at every cluster box in parentheses. Two or three upper lines at the box are related to a partial taxonomy composition of a subcluster. Further lines mention growth temperature contents. Bottom line indicates an average $A + T$-content and its standard deviation in percents. A schematic profile of a subcluster is placed at the bottom-right corner of a box.

4. Discussion and conclusions

In this study we used a DNA curvature excess profile technique to reduce a comprehensively big text file (genome) to a numerical vector consisting of 801 real positive numbers smaller than 0.5. Such 205 vectors in a multidimensional Euclidean space were used for further data clustering based on two very widespread methods: *k*-means and PAM. The results obtained by *k*-means algorithm application seem to possess better biological relevance. We applied *k*-means algorithm to cluster genomes using curvature excess distributions upstream of the starts of genes (putative promoter regions). We found that the main factors influencing curvature distribution in promoter regions of the prokaryotes are, in order of importance:

Table 3
Correlation between taxonomy and G5MCC

Phylum	Class	Genomes with G5MCC	Genomes without G5MCC
Actinobacteria	Actinobacteria	12	4
Chlamydiae	Chlamydiae	1	6
Cyanobacteria		2	5
Firmicutes	Bacilli	22	8
Firmicutes	Mollicutes	2	8
Proteobacteria	Alphaproteobacteria	15	5
Proteobacteria	Betaproteobacteria	10	0
Proteobacteria	Gammaproteobacteria	37	7
Proteobacteria	Deltaproteobacteria	4	2
Proteobacteria	Epsilonproteobacteria	6	0
Spirochaetes	Spirochaetes	1	7

- optimal growth temperature;
- genome size;
- $A + T$ composition.

The possible combinations among these factors influencing curvature excess can explain, for example, the homogeneous distribution of mesophilic genomes along the clusters. The absence of excessive curvature in almost all thermophiles and hyperthermophiles brings them all together in a mutual cluster using clustering based on promoter or terminator profiles (data not shown); while the majority of the mesophilic *AT*-rich genomes were located in other clusters.

In any analysis, including comparative genomics, correlation among different factors should be taken into account [9,8]. We aware of the fact that among three main factors: optimal growth temperature, genome size, AT content, — the latter two are related. Really big genomes tend to be GC rich, and the small genomes are AT rich. This does not corroborate our conclusion that Cluster 1 (the cluster containing genomes with the highest curvature excess values in promoter regions) contains exclusively mesophilic prokaryotes that have big AT rich genomes.

Evolution of proteomes was studied in [29,20]. Their analyses revealed striking discrimination between mesophiles and hyperthemophiles, following amino acid usage. There is no influence of amino acid usage on curvature excess; however, both analyses found that optimal growth temperature is the main discriminative factor.

Clustering based on DNA curvature distributions in coding regions was performed for the first time, as far as we know. The cluster analysis provides some insight into the phenomenon of surprisingly high DNA curvature located immediately after the start of the gene. Why would evolution frequently keep an extensive curvature at the 5-ends of many protein-coding sequences? Why is the phenomenon of G5MCC so typical for many representatives of proteobacteria but seemingly a rather rare episode for Chlamydiae or Spirochaetes? At the moment, we do not have any solid answers to the second question. However, we can suggest some answers to the first question.

The first hypothesis that should be investigated (study in progress) suggests that observed excess of the DNA curvature mirrors alternation of hydrophobic and hydrophilic amino acids in α-helices [36,10]. To support or reject this hypothesis we would check how frequently the α-helix structure appears at the very beginning of a gene.

Another (but non-contradictive) explanation suggests that 5-end of a gene may be a frequent site of DNA-protein interactions, involving an unidentified histon-like protein that preferably binds to curved DNA. Some indirect evidence is brought in following lines. As it is well known, many bacterial genes are grouped in operons. Operons frequently contain additional regulatory elements as attenuation of transcription [34] (pp. 128–136). Some operons have a leader peptide with no apparent function but contain an effective shifting terminator. We have recently found that terminators of many prokaryotes have significant downstream curvature [18,11]. We performed a preliminary analysis to get a clearer picture of whether a curvature peak typical for all G5MCC serves as part of a promoter sequence or belongs to another regulation area. For this purpose we extracted operon sequences with a leader peptide from *E. coli* and constructed curvature profiles for these regions. The results have shown that many of these operons possess curved DNA sequences around the starts of the first structural genes (data not shown). Since these curved sequences are separated from the promoter by a leader peptide, we assume that this phenomenon cannot be explained by the curved binding elements frequently present in promoter sites of prokaryotic genomes; on the contrary, we can speculate that they are involved in attenuation of transcription. The role of curved DNA in a termination process is still not clear. Our hypothesis is that curved elements assist in an effective termination by slowing down RNAp, either indirectly through an additional factor, such as histone-like nucleoid structuring protein (H-NS), or directly through the curved sequences. Future biochemistry analysis is needed to clarify biological functions of curved DNA in terminator and anti-terminator regions. While we found that the main factor influencing inclination in curvature distribution in promoter regions of the prokaryotes is optimal growth temperature, the sharp maximum of curvature excess immediately at the 5-end of coding sequences (G5MCC) is strongly related to the taxonomy of prokaryotes. Considering the taxonomy description of G5MCC, the picture is quite homogeneous: the absolute majority of the phylum Proteobacteria is located in G5MCC; practically all big Firmicutes are in G5MCC as well (see Fig. 5, Table 3, and for more details Supplementary Table). Interestingly, $A + T$ composition of G5MCC is rather heterogeneous.

Appendix. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.dam.2008.06.049.

References

[1] S.D. Bentley, J. Parkhill, Comparative genomic structure of prokaryotes, Ann. Rev. Genet. 38 (2004) 771.
[2] A. Bolshoy, P. McNamara, R.E. Harrington, E.N. Trifonov, Curved DNA without A-A - experimental estimation of all 16 DNA wedge angles, Proc. Natl. Acad. Sci. USA 88 (6) (1991) 2312.
[3] A. Bolshoy, E. Nevo, Ecologic genomics of DNA: Upstream bending in prokaryotic promoters, Genome Res. 10 (8) (2000) 1185.
[4] S. Diekmann, J.C. Wang, On the sequence determinants and flexibility of the kinetoplast DNA fragment with abnormal gel-electrophoretic mobilities, J. Mol. Biol. 186 (1) (1985) 1.
[5] E.W. Forgy, Cluster analysis of multivariate data — efficiency vs interpretability of classifications, Biometrics 21 (3) (1965) 768.
[6] C. Fraley, A.E. Raftery, How many clusters? which clustering method? Answers via model-based cluster analysis, Comput. J. 41 (8) (1998) 578.
[7] J. Griffith, M. Bleyman, C.A. Rauch, P.A. Kitchin, P.T. Englund, Visualization of the bent helix in kinetoplast DNA by electron-microscopy, Cell 46 (5) (1986) 717.
[8] P.F. Hallin, T.T. Binnewies, D.W. Ussery, Genome update: Chromosome atlases, Microbiology-Sgm 150 (2004) 3091.
[9] T. Hallin, P.F. Coenye, T.T. Binnewies, H. Jarmer, H.H. Saerfeldt, D.W. Ussery, Genome update: Correlation of bacterial genomic properties, Microbiology-Sgm 150 (2004) 3899.
[10] H. Herzel, O. Weiss, E.N. Trifonov, 10–11 bp periodicities in complete genomes reflect protein structure and DNA folding, Bioinformatics 15 (3) (1999) 187.
[11] S. Hosid, A. Bolshoy, New elements of the termination of transcription in prokaryotes, Biomol. Struct. Dyn. 22 (3) (2004) 347.
[12] R. Jauregui, C. Abreu-Goodger, G. Moreno-Hagelsieb, J. Collado-Vides, E. Merino, Conservation of DNA curvature signals in regulatory regions of prokaryotic genes, Nucleic Acids Res. 31 (2003) 6770.
[13] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, Wiley and Sons, New York, 1990.
[14] L. Klasson, S.G.E. Andersson, Evolution of minimal-gene-sets in host-dependent bacteria, Trends Microbiol. 12 (1) (2004) 37.
[15] J. Kogan, C. Nicholas, V. Volkovich, Text mining with hybrid clustering schemes, in: M.W. Berry and W.M. Pottenger (Eds.), Proceedings of the Workshop on Text Mining, 2003, held in conjunction with the Third SIAM International Conference on Data Mining, p. 516.
[16] L. Kozobay-Avraham, A. Bolshoy, Z. Volkovich, On prokaryotes' clustering based on curvature distribution, in: Advances in Web Intelligence and Data Mining, in: M. Last, P.S. Szczepaniak, Z. Volkovich, A. Kandel (Eds.), Studies in Computational Intelligence, vol. 23, Springer-Verlag, NY, 2006, p. 275.
[17] L. Kozobay-Avraham, S. Hosid, A. Bolshoy, Curvature distribution in prokaryotic genomes, In Silico Biol. 4 (3) (2004) 361.
[18] L. Kozobay-Avraham, S. Hosid, A. Bolshoy, Involvement of DNA curvature in intergenic regions of prokaryotes, Nucleic Acids Res. 34 (8) (2006) 2316.
[19] W.J. Krzanowski, Y.T. Lai, A criterion for determining the number of groups in a data set using sum-of-squares clustering, Biometrics 44 (1) (1988) 23.
[20] J.R. Lobry, A. Necsulea, Synonymous codon usage and its potential link with optimal growth temperature in prokaryotes, Gene 385 (2006) 128.
[21] J. Mardia, K. Kent, J. Bibby, Multivariate Analysis, Academic Press, San Diego, 1979.
[22] J.C. Marini, S.D. Levene, D.M. Crothers, P.T. Englund, Bent helical structure in kinetoplast DNA, Proc. Natl. Acad. Sci. USA 79 (24) (1982) 7664.
[23] N. Olivares-Zavaleta, R. Jauregui, E. Merino, Genome analysis of Escherichia coli promoter sequences evidences that DNA static curvature plays a more important role in gene transcription than has previously been anticipated, Genomics 87 (2006) 329.
[24] A.G. Pedersen, L.J. Jensen, S. Brunak, H.H. Staerfeldt, D.W. Ussery, A DNA structural atlas for *Escherichia coli*, Mol. Biol. 299 (2000) 907.
[25] J. Perez-Martin, F. Rojo, V. L d, Promoters responsive to DNA bending: A common theme in prokaryotic gene expression, Microbiol. Rev. 58 (1994) 268.
[26] E.S. Shpigelman, E.N. Trifonov, A. Bolshoy, Curvature—software for the analysis of curved DNA, Comput. Appl. Biosci. 9 (4) (1993) 435.
[27] C.A. Sugar, G.M. James, Finding the number of clusters in a dataset: An information-theoretic approach, J. Amer. Statist. Assoc. 98 (463) (2003) 750.
[28] M. Teboulle, P. Berkhin, I. Dhillon, Y. Guan, J. Kogan, Clustering with entropy-like *k*-means algorithms, in: J. Kogan, C. Nicholas, M. Teboulle (Eds.), Grouping Multidimensional Data: Recent Advances in Clustering, Springer-Verlag, NY, 2006, p. 127.
[29] F. Tekaia, E. Yeramian, Evolution of proteomes: Fundamental signatures and global trends in amino acid compositions, BMC Genomics 7 (2006) 307.
[30] E.N. Trifonov, L.E. Ulanovsky, Inherently curved DNA and its structural elements, in: Wells R.D., Harvey S.C. (Eds.), Unusual DNA Structures, Springer-Verlag, Berlin, 1987, p. 173.
[31] L. Ulanovsky, M. Bodner, E.N. Trifonov, M. Choder, Curved DNA - design, synthesis, and circularization, Proc. Natl. Acad. Sci. USA 83 (4) (1986) 862.
[32] D.W. Ussery, P.F. Hallin, Genome update: length distributions of sequenced prokaryotic genomes, Microbiology-Sgm 150 (2004) 513.
[33] D.W. Ussery, N. Tindbaek, P.F. Hallin, Genome update: Promoter profiles, Microbiology-Sgm 150 (2004) 2791.
[34] R. Wagner, Transcription Regulation in Prokaryotes, Oxford University Press, USA, 2000.
[35] H.M. Wu, D.M. Crothers, The locus of sequence-directed and protein induced DNA bending, Nature 308 (5959) (1984) 509.
[36] V.B. Zhurkin, Periodicity in DNA primary structure is defined by secondary structure of the coded protein, Nucleic Acids Res. 9 (8) (1981) 1963.